

# RUMD: A general purpose molecular dynamics package optimized to utilize GPU hardware down to a few thousand particles

Nicholas P. Bailey,<sup>1,\*</sup> Trond S. Ingebrigtsen,<sup>1</sup> Jesper Schmidt Hansen,<sup>1</sup> Arno A. Veldhorst,<sup>1</sup> Lasse Bøhling,<sup>1</sup> Claire A. Lemarchand,<sup>1</sup> Andreas E. Olsen,<sup>1</sup> Andreas K. Bacher,<sup>1</sup> Lorenzo Costigliola,<sup>1</sup> Ulf R. Pedersen,<sup>1</sup> Heine Larsen,<sup>1</sup> Jeppe C. Dyre,<sup>1</sup> and Thomas B. Schrøder<sup>1,†</sup>

<sup>1</sup>*“Glass and Time”, IMFUFA, Dept. of Science and Environment,  
Roskilde University, P.O. Box 260, DK-4000 Roskilde, Denmark*

(Dated: September 9, 2016)

RUMD is a general purpose, high-performance molecular dynamics (MD) simulation package running on graphical processing units (GPU’s). RUMD addresses the challenge of utilizing the many-core nature of modern GPU hardware when simulating small to medium system sizes (roughly from a few thousand up to hundred thousand particles). It has a performance that is comparable to other GPU-MD codes at large system sizes and substantially better at smaller sizes. RUMD is open-source and consists of a library written in C++ and the CUDA extension to C, an easy-to-use Python interface, and a set of tools for set-up and post-simulation data analysis. The paper describes RUMD’s main features, optimizations and performance benchmarks.

## I. INTRODUCTION

This paper describes Roskilde University Molecular Dynamics (RUMD), a Graphical Processing Unit (GPU)-based molecular dynamics (MD) code developed to achieve good performance at small and medium system sizes, while remaining competitive with other GPU-MD codes at large sizes. The attention paid to small sizes distinguishes RUMD from many other GPU-MD codes. It has been in development since 2009, and available as open-source software<sup>1</sup>, since 2011. The newest version 3.1, was released May 2016.

The rise of GPU-based computation has been discussed by various authors<sup>2–8</sup>. Several groups have developed molecular dynamics codes based on GPUs from scratch or incorporated GPU-acceleration into existing projects. Examples of the former include HOOMD-Blue<sup>9–12</sup>, ACEMD<sup>13</sup>, OpenMM<sup>14,15</sup> and HAL’s MD<sup>16</sup> while the latter include NAMD<sup>17</sup>, LAMMPS<sup>18</sup>, AMBER<sup>19</sup> and Gromacs<sup>20</sup>. Other works involving GPU-based MD codes, going back to 2007, can be found in Refs. 21–32. We omit a detailed exposition of GPU programming basics here. For a good overview of massive multi-threading using CUDA see the relevant section in the article by Anderson *et al.*<sup>9</sup>. For further information the reader can consult the book by Kirk and Hwu<sup>33</sup> as well as the CUDA programming guide<sup>34</sup>. A technical overview of the Tesla architecture, which marks the first major development of GPUs for scientific computing by NVIDIA, can be found in Ref. 35. The more advanced Fermi architecture is documented in Ref. 36. The most recent architectures Kepler (2012) and Maxwell (2014) are described in Ref. 37 and 38 respectively.

The large computational power of modern GPUs comes primarily from the large number of hardware cores, each executing a number of software threads. As an example, the GeForce Gtx780Ti card has 2880 cores and a theoretical single-precision peak-performance of 5.0 TFlops ( $5 \times 10^{12}$  floating point operations per second). A key element to achieve good performance from a GPU is that the number of active software threads should be much larger than the number of hardware cores in order to hide latency of memory access.

This makes it a challenge to utilize the GPU hardware when the number of particles  $N$  is relatively small ( $N \sim 10^3 - 10^4$ ). The obvious choice for parallelization, namely having one thread compute the forces for one particle, is clearly not efficient when the optimal number of threads exceeds the number of particles. There are three reasons to focus on utilizing the GPU hardware even at small system sizes; i) If one is interested in investigating long time scales rather than large systems. This is the case, for example, in the field of viscous liquid dynamics, where a system size of  $10^4$  particles is considered large, but the interest is in studying as long time scales as possible. Note that finite size effects are relatively limited in these systems; for example Karmakar, Dasgupta, and Sastry<sup>39</sup> found convergence of diffusivity and relaxation time for a standard glass-forming model liquid already at  $N=1000$ . ii) As a building block for multi-GPU simulations (RUMD currently uses one GPU per simulation). If one wants to simulate, say,  $10^5$  particles using 10 GPU’s, the single-GPU performance obviously needs to be good for  $10^4$  particles. iii) For the foreseeable future much of the development in GPU and other many-core hardware will probably be in increasing the number of physical cores, much more than increasing the computational power of the individual core. Thus, what might today be considered a large system, might in the future be considered a small/medium sized system where special care needs to be taken to utilize the GPU hardware. To optimize the use of the hardware RUMD allows multiple threads per particle; this approach has also been considered in two recent publications<sup>40,41</sup>.

The paper is organized as follows. Section II contains a brief overview of RUMD’s features. The main part of the paper focuses on the methods used for calculating the non-bonding pair interactions and the generation of the neighborlist. These are the most computationally demanding parts of an MD simulation and where our code distinguishes itself from most other GPU-MD codes. Section III discusses the challenges of utilizing the GPU hardware at small system sizes, and section IV gives an overview of the optimization strategies employed in RUMD. Section V describes the calculation of non-bonding pair-forces, while sections VI and VII

describes two different methods for generating the neighbor-list. Section VIII provides benchmarks of RUMD in comparison to three different GPU extensions of LAMMPS<sup>18</sup>, as well as an analysis of the effect of the different optimizations employed in RUMD. Section IX describes RUMD’s performance for electrostatic (Coulomb) interactions. Section X provides a short summary.

## II. RUMD: FEATURES

RUMD is a general purpose molecular dynamics code. Below we list its main features; for more information please see the tutorial and user manual included with the software and available from the project’s website `rumd.org`.

**Python interface:** Users control the software via a Python interface which allows simulations of considerable complexity to be implemented straightforwardly. An example of a simple user Python-script is given in Fig. 1.

**Pair potentials:** 12-6 Lennard-Jones, generalized Lennard-Jones, inverse power law, Gaussian core, Buckingham, Dzugotov, Girifalco, Yukawa, and more. New pair potentials are easily added, as described in the tutorial. Three different “cutoff methods” for truncating the pair potential are provided: simple truncation with no shift; truncation plus shift of the potential energy to ensure continuity; and truncation plus shift of the pair force<sup>42</sup> to ensure its continuity (this corresponds to adding a linear term in the potential).

**Other interactions:** Intramolecular interactions including constraints, bond-stretching forces, angular forces and dihedral forces.

**Integrators:** NVE (Verlet/Leap-frog), NVT (Nosé-Hoover), NPT<sup>43</sup>, NVU (geodesics on the constant potential energy surface)<sup>44,45</sup>. Couette shear flow using the SLLOD equations of motion and Lees-Edwards boundary conditions.

**File formats:** configurations are stored in xyz format with extensions, compressed using gzip; data can be saved logarithmically in time for efficient use of disk space while allowing the study of a large range of time scales in a single simulation; molecular structure (bonds, angles and dihedrals) is specified in separate topology files. Tools for creating initial configurations and topology files are provided.

**Analysis tools:** Basic statistics of energy, pressure, etc. for thermodynamics. Measures of structure; radial distribution function, static structure factor, radius of gyration, mean-square end-to-end distance. Measures of dynamics; mean-square displacement, incoherent intermediate scattering function, non-Gaussian parameter, end-to-end vector autocorrelation function, Rouse-mode autocorrelation function. New analysis tools are continuously being added. Analysis tools work on data stored

during simulations and can thus be applied at the end of (or during) a simulation run. In addition the user can define customized on-the-fly analysis tools written in Python.

**Auto-tuner:** A script for optimizing internal parameters—specifically, the choice of algorithm for generating the neighbor list, the neighbor-list skin size, and the way the generation of the neighbor list and the calculation of non-bonding forces are distributed among the GPU threads.

RUMD is mostly implemented in single precision. This leads to a drift in the total energy when running long constant-energy (NVE) simulations, but is not an issue for NVT and NPT simulations where a thermostat is applied. RUMD can be made fully double precision by a search and replace in the source code - we are planning to implement a more elegant way for the user to choose between single and double precision. RUMD uses a single GPU per simulation; support for multiple GPU simulations is planned for future development.

## III. THE PROBLEM OF UTILIZING THE DEVICE AT SMALL SYSTEM SIZES

Consider NVIDIA’s Kepler GK110 architecture that appeared in 2013. One of the Kepler design goals was power efficiency, which was partly achieved by increasing the number of cores while decreasing the clock speed compared to the previous Fermi architecture. Thus each streaming multiprocessor (of type SMX) has 192 cores, and the GPU has up to 15 SMX units. The GTX 780Ti card contains the maximum 15 SMX units, giving 2880 cores. Furthermore, the CUDA model requires a much greater number of threads to be active, in order to hide memory access latency. This poses a challenge when small systems of the order of thousands of particles are concerned. Logically, in order to use as many threads as possible, one must therefore have multiple threads computing the force on one particle.

Having multiple threads per particles entails some overhead, in particular the summing of the force contributions over the threads allocated to a given particle. This means that as the system size increases, it becomes less useful to have more than one thread per particle. We control this by the parameter  $t_p$  (threads per particle, denoted `TPerPart` in the code), and let the auto-tuner pick the optimal value for a given simulation. The optimal value of  $t_p$  depends primarily on the number of particles, but also on density and the range of the potential. We use a separate kernel involving a single thread per particle for larger sizes (see Fig. 2); this is faster than setting  $t_p = 1$  in the general kernel.

Rovigatti *et al.* have recently discussed the possible advantages of “vertex-based” (atom-decomposition<sup>46</sup>, one thread per particle) versus “edge-based” (force-decomposition<sup>46</sup>, one thread per interaction) parallelism<sup>47</sup>. Our approach includes the former, and a range of intermediate cases, while not taking it to the extreme of one thread per interaction.

```

# Import RUMD
from rumd import *
from rumd.Simulation import Simulation

# Create a simulation object, and import an initial configuration.
sim = Simulation("start.xyz.gz")

# Create a pair potential and associate it with the simulation object
pot = Pot_LJ_12_6(cutoff_method=ShiftedForce)
pot.SetParams(0, 0, Sigma=1.0, Epsilon=1.0, Rcut=2.5)
sim.SetPotential(pot)

# Create an integrator and associate it with the simulation object
itg = IntegratorNVT(timeStep=0.004, targetTemperature=1.0)
sim.SetIntegrator(itg)

# Run a simulation. Data are saved on disk and can be analyzed by a number of tools
sim.Run(1000000)

```

FIG. 1: Script showing the python code needed to run a very simple simulation, in this case a single-component Lennard-Jones fluid simulated at constant temperature 1.0 for one million time steps of size 0.004 in Lennard-Jones units. The number of particles and the density is determined by the initial configuration contained in the file start.xyz.gz

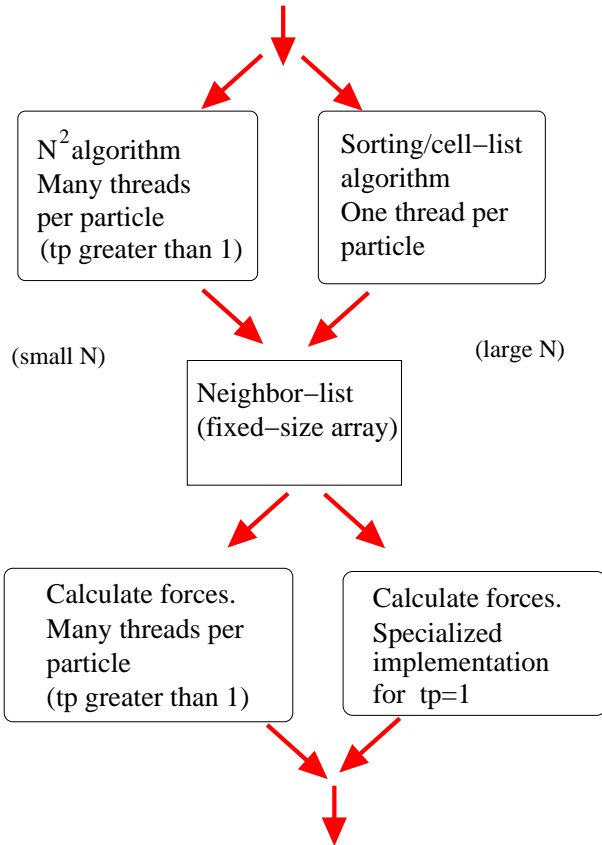


FIG. 2: Schematic diagram representation of the two algorithms for neighbor-list generation, and the force calculation algorithm. The latter uses multiple threads per particle (tp), but an implementation also exists for the special case tp=1.

#### IV. OVERVIEW OF OPTIMIZATION STRATEGIES USED IN RUMD

As in any general purpose MD software some kind of data structure to keep track of neighbors for the non-bonding pair interactions is necessary to reduce the complexity of the force calculation from  $O(N^2)$  to  $O(N)$ . We use a classical Verlet-type neighbor list, stored as 2-dimensional fixed-size array of size  $Nn_{\max}$  where  $n_{\max}$  is the assumed maximum number of neighbors per particle. If this happens to be exceeded the neighbor-list is automatically re-allocated with doubled capacity. For smaller systems we set  $n_{\max} = N$  from the start to avoid the overhead of checking for overflow. Neighbors within  $r_c + s$  are listed, where  $r_c$  is the maximum cut-off associated with the potential, and  $s$  is the extra skin included so that the neighbor-list does not need to be rebuilt every step. The optimal value of the skin is determined by the auto-tuner.

We now describe the methods employed in the calculation of short-range non-bonding forces and the generation of the neighbor-list. The main four optimizations are as follows:

1. Multiple threads per particle ( $t_p \geq 1$ ) in force calculation and neighbor-list generation. The auto-tuner chooses the best value for  $t_p$ .
2. Two methods for rebuilding the neighbor-list:  $O(N^2)$  method ( $t_p \geq 1$ ) for small system sizes, and an  $O(N)$  method ( $t_p = 1$ ) for larger sizes. The auto-tuner picks the best method.
3. Use of the so-called “read only data-cache” for reading positions (for devices of compute capability at least 3.5 this can be done straightforwardly via the function `__ldg()`).
4. Use of pre-fetching when reading from the neighbor-list to compensate for memory access latency.

```

__global__ void Calcf_NBL_tp( ... )
[ Declare shared memory ]
float4 my_f = {0.0f, 0.0f, 0.0f, 0.0f}; // Initialize the force of this thread
float4 my_r = LOAD(r[MyGP]); // Read position of this particle
int type_i = __float_as_int(my_r.w); // Type of this particle
[ Read information on the simulation box from device memory ]
[ Copy potential parameters to shared memory ]
__syncthreads(); // Parameters loaded to shared memory before proceeding

int nb, my_num_nbrs = num_nbrs[MyGP]; // Read number of neighbors
nb_prefetch = nbl[nvp*MyT + MyGP]; // Read index of first neighbor
for (int i=MyT; i<my_num_nbrs; i+=TPerPart) { // Loop over neighbors
    nb = nb_prefetch;
    if(i+TPerPart < my_num_nbrs)
        nb_prefetch = nbl[nvp*(i+TPerPart) + MyGP];
    float4 r_i = LOAD(r[nb]); // Read position of neighbor
    int type_i = __float_as_int(r_i.w); // Type of neighbor
    // Add contribution from this pair to my_f:
    fij( potential, my_r, r_i, &my_f, [parameters, simulation box] );
}
__syncthreads();

// Now use the shared memory for summing force contributions:
s_r[MyP+MyT*PPerBlock] = my_f;
__syncthreads();
// Sum over threads associated with the same particle:
if( MyT == 0 ) {
    for( int i=1; i < TPerPart; i++ ) my_f += s_r[MyP + i*PPerBlock];
    my_f.w *= 0.5f; // Compensate for double counting of potential energy
    // Write result to device memory
    if(initialize) // Flag (templated) for whether to initialize forces
        f[MyGP] = my_f; // (can be false when multiple potentials present)
    else {
        atomicAdd(&(f[MyGP].x), my_f.x);
        [also y, z, w]
    }
}
}
}

```

FIG. 3: Kernel calculating forces on particles given the neighbor-list (nbl) shown in the simplest version where only forces and potential energy of each particle are computed. For a given particle each of  $t_p$  threads ( $MyT = 0, 1, \dots, t_p - 1$ ) computes part of the total force, which is summed up at the end. The function `fij` (not shown) adds an individual pair contribution to the current thread's force (`my_f`). Note the use of `__syncthreads` to synchronize threads within a thread-block. This is to ensure that all data are available in shared memory before any thread reads from it (first and third occurrences) or that all threads are done with the data in shared memory before it is used for other data (second occurrence). `LOAD()` is a macro that reads from device memory via the read only data-cache using `__ldg()` on cards where this is available.

TABLE I: Short-hand notation for common quantities used in CUDA-kernels.

quantity	name in kernel	CUDA variable
Number of thread-blocks	NumBlocks	gridDim.x
Particles per block ( $p_b$ )	PPerBlock	blockDim.x
Threads per particle ( $t_p$ )	TPerPart	blockDim.y
Particle index within block	MyP	threadIdx.x
Thread index w.r.t. particle	MyT	threadIdx.y
Index of thread-block	MyB	blockIdx.x
Global index of particle	MyGP	MyP+MyB*PPerBlock

## V. FORCE CALCULATION

The force calculation kernel (routine executed on the GPU) is shown in Fig. 3. Short-hand notation for common quantities used in this and the following CUDA-kernels are given in Table I. The force kernel uses in general  $t_p \geq 1$ , although a separate implementation for  $t_p = 1$  (not shown) was made because at large sizes it is no longer beneficial to have more

than one thread per particle (there are many threads anyway), and the overhead associated with summing over threads is noticeable. The neighbor-list is arranged in column-major order, i.e., the first neighbors of all particles are consecutive in memory, then the second neighbors, etc. This allows for efficient (coalesced) memory access.

Note the use of pre-fetching when reading from the neighbor-list; while the force contribution of neighbor  $i$  is computed, the index of neighbor  $i + 1$  is being read from the neighbor list.

Within the kernel a call is made to a function `fij` (not shown), which calculates the contribution to the pair force on the current particle from a neighbor particle. `fij` itself calls a function `ComputeInteraction` which is unique to each type of pair potential, and selected via templating. Templating is used so that it is known when compiling `fij` which potential, and thus which `ComputeInteraction`, is to be called. Templating is also used for some of the other user-chosen variables, including the type of boundary conditions (represented by a `SimulationBox` class) and the cutoff-method. This means that the force calculation kernel is com-



```

__global__ void calculateNBL_N2( ... ) {
  const unsigned int tid = MyP + MyT*PPerBlock; // Thread-index within block
  [ Declare shared memory: s_r, s_Count, s_cut_skin2 ]

  if (updateRequired[0]) {
    if (MyT==0) s_Count[MyP]=0; // Count of neighbors for this particle
    [ Copy cut-offs plus skin squared to shared memory ]
    float4 my_r = r[MyGP]; // Position of this particle

    // Loop over blocks of particles
    for (FirstGP=0; FirstGP<numParticles; FirstGP+=TPerPart*PPerBlock) {

      // Read particle positions in block into shared memory
      if (FirstGP + tid<numParticles) s_r[tid] = r[FirstGP + tid];
      __syncthreads(); // Shared data in s_r ready

      // Loop over particles in block
      for (int i=0; i<PPerBlock*TPerPart; i+=TPerPart) {
        OtherP = i + MyT; OtherGP = FirstGP + OtherP;
        if (MyGP<numParticles && MyGP!=OtherGP && OtherGP < numParticles) {
          float4 r_i = s_r[OtherP]; // Position of other particles
          [ Read squared cutoff distance from shared memory based on types ]
          [ Calculate squared distance dist2 ]
          if (dist2 < RcutSk2) {
            // Atomically increment counter for this particle:
            unsigned int nextNbrIdx = atomicInc(&s_Count[MyP], numParticles);
            [ If space insert index into NB-list at position nextNbrIdx ]
          } // if(dist2 < RcutSk2)
        } // if (MyGP < numParticles)
      } // for(int i ... )
      __syncthreads();
    } // for (int firstGP ... )
    __syncthreads();

    if (MyT == 0) {
      [ Store this particles number of neighbors ]
      [ Store its position so can check when rebuild needed ]
      [ Detect whether ran out of space and set flag to inform host ]
      if (MyP == 0) atomicDec(&(updateRequired[0]), NumBlocks);
    } // if (MyT == 0 ... )
  } // if(update_required[0])
}

```

FIG. 4: Kernel for order- $N^2$  neighbor-list generation. Note that because the number of particles is not in general a multiple of  $p_b$ , there are some threads in the last block which should not do anything, hence statements such as `if (MyGP<numParticles)`.

piled for all possible combinations of these parameters, and the user can choose the appropriate one at run time. The code for the conditional statements which allows this is tedious, but can be generated automatically by a Python script. The main disadvantage of using templating is that it increases the compile time considerably.

## VI. NEIGHBOR-LIST GENERATION: ORDER- $N^2$

This neighbor-list generating algorithm has  $O(N^2)$  complexity and is thus suitable only for small system sizes. In a serial code there would be a double loop; in a parallel code one loop (over particles whose neighbors are to be found) are handled completely by parallelization. Part of the loop over “other” particles is handled by looping over  $t_p$ -sized groups, while parallelization (the  $t_p$  threads for that particle) accounts for looping within these groups (we do not make use of Newton’s third law). Shared memory is used to reduce the amount of reads from device memory; in a straight-forward implementation without shared memory, a total of  $N^2$  reads of particle positions is necessary. By using a block-wise reading

into the shared memory, this is reduced to  $N^2/p_b$ , where  $p_b$  is the number of particles in a block (denoted `PPerBlock` in the code). From this consideration  $p_b$  should be as large as possible, but on the other hand a too large  $p_b$  value would mean that the number of blocks ( $\approx N/p_b$ ) becomes too small to occupy the number of SMX multiprocessors. RUMD uses the auto-tuner to pick the best value for  $p_b$ .

The kernel uses  $t_p$  threads for a given particle to search for neighbors. This means that we have to deal with the situation that several or all of them find a neighbor at the same time, and the writing to the neighbor list should be performed without race-conditions. This is achieved by a so-called *atomic* operation. When several threads perform an atomic operation on the same variable, all the operations are guaranteed to be performed in (an unspecified) sequential order. Here we use the atomic increment function, `atomicInc()`, which ensures that the number of neighbors is counted correctly. When a thread is calling `atomicInc()`, the function returns the value that the variable had *before* the increment of the given thread is performed. This is here used to specify a unique position in the neighbor list (`nextNbrIdx`).

The information about whether the neighbor-list needs to be

```

__global__ void calculateNBL_CellsSorted( ... ) {
    gtid = blockIdx.x*blockDim.x + threadIdx.x; Count = 0;
    [ Declare shared memory: s_r, s_cut_skin2 ]
    [ Copy cut-offs plus skin squared to shared memory ]
    __syncthreads();
    if (gtid < numParticles) {
        float4 my_r = r[gtid];
        int3 my_CellCoordinates = calculateCellCoordinates(my_r, ...);
        int3 OtherCellCoordinates;

        // Loop over neighboring cells, applying periodic boundary conditions
        for (int dZ=-2; dZ<=2; dZ++) {
            OtherCellCoordinates.z = (my_CellCoordinates.z + dZ + num_cells_vec.z)%num_cells_vec.z;
            for (int dY=-2; dY<=2; dY++) {
                OtherCellCoordinates.y = (my_CellCoordinates.y + dY + num_cells_vec.y)%num_cells_vec.y;
                for (int dX=-2; dX<=2; dX++) {
                    OtherCellCoordinates.x = (my_CellCoordinates.x + dX + num_cells_vec.x)%num_cells_vec.x;

                    // Loop over particles in cell
                    int otherCellIndex = calculateCellIndex(OtherCellCoordinates, num_cells_vec);
                    int Start = cellStart[otherCellIndex];
                    int End = cellEnd[otherCellIndex];
                    for (int OtherP=Start; OtherP<=End; OtherP++) {
                        if (gtid != OtherP) {
                            float4 r_i = LOAD(r[OtherP]);
                            [ Read squared cutoff distance from shared memory based on types ]
                            [ Calculate squared distance dist2 ]
                            if (dist2 < RcutoffSk2)
                                [ If space insert index into NB-list and increment Count, else break ]
                        }
                    } // end for (int OtherP...)
                }
            }
        } // end for (int dZ ...)
        [ Store this particles number of neighbors ]
        [ Store its position so can check when rebuild needed ]
        [ Detect whether ran out of space and set flag to inform host ]
        if ( gtid==0 ) updateRequired[0] = 0;
    } // if (gtid < numParticles)
}

```

FIG. 5: Kernel for order- $N$  neighbor-list generation. `calculateCellCoordinates(...)` calculates the coordinates of the cell that a particle belongs to. `calculateCellIndex(...)` calculates the index of a cell given its coordinates. The arrays `cellStart` and `cellEnd` contain the indices of the first and last particles, respectively, associated with a given cell.

rebuild is on the device, generated by a different kernel. The kernel in Fig. 4 is called at every timestep, and then checks via `if(updateRequired)` whether there is anything to be done. This is faster than copying the value of a flag to the host and having the host decide whether to launch the rebuild-kernel. `updateRequired` is initially equal to the number of thread-blocks. One thread from each block decrements it with an atomic operation (`atomicDec()`) when it (its thread-block) is done, so that when all blocks are finished it is zero. At the next time step, assuming no particles have moved more than half the skin distance, `updateRequired` will still be zero and therefore the kernels will immediately exit. Using an atomic operation to decrement `updateRequired` is necessary because the thread-blocks execute asynchronously, so none of them knows when/whether the others are finished, or even started—any unfinished blocks need to be able to see a non-zero value of the counter.

The above means that for small systems the simulations are performed entirely on the GPU without any communication with the CPU (except when output is required). Avoiding the overhead associated with communication between the GPU and CPU is important for the performance at small system sizes.

## VII. NEIGHBOR-LIST GENERATION: ORDER- $N$

The order- $N$  algorithm is based on a cell-index method<sup>48</sup> and involves (1) dividing the simulation box into rectangular spatial cells whose size is related to the potential cutoff; (2) associating particles with the appropriate cell based on the coordinates; (3) sorting the particles according to cell-index and rearranging all particle data to the sorted order (this can be done quickly with the Thrust library<sup>49</sup>). The advantage of rearranging the particle data to the sorted order is two fold; i) the information about which particles are in a given cell can be stored simply as two integers indicating the first (`cellStart`) and the last particle (`cellEnd`); ii) better performance of the data-cache when reading the particle information both in the neighbor list creation in the force calculation.

The kernel in Fig. 5 is called after steps (1) to (3) have been carried out via a series of small kernels and Thrust operations. It involves, for a given particle, identifying its cell coordinates and looping over neighboring cells in three dimensions to find neighbors. We have chosen cell lengths in each direction to be of order (not less than)  $(r_c + s)/2$ , where  $s$  is the neighbor-list skin. This means that the loop extends to plus/minus two cells in each direction, or 125 cells altogether. Such a choice of

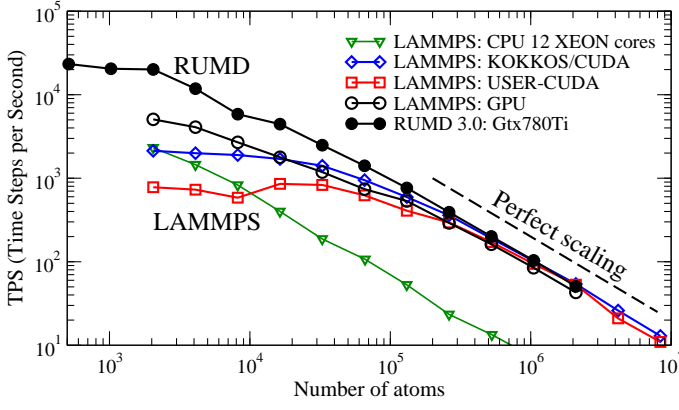


FIG. 6: The LAMMPS benchmark: a melting FCC crystal is simulated at constant energy. The vertical axis shows the number of simulated time steps per second of wall-clock time. At large system sizes all codes follows the ideal  $1/N$  scaling, and the GPU's are 10-20 faster than LAMMPS running on 12 xeon cores. For RUMD good scaling is maintained down to quite small systems  $N \sim 2000$ , and at small system sizes RUMD is thus considerably faster than the three GPU versions of LAMMPS.

cell length means one searches a volume 58%  $[(125/8)/27]$  of that searched when using cells of length  $r_c + s$ . This kernel is called with one thread per particle, since that is generally most efficient at larger sizes, which is also when the linear method of neighbor-list generation becomes relevant. It is conceivable that some gain at intermediate sizes could be achieved by implementing a  $t_p > 1$  version of the kernel, but this has not been tried yet.

In this case the information about whether to rebuild the neighbor-list must be communicated to the host because several kernels and Thrust functions must be called (the use of Dynamic Parallelism, available since CUDA 5.0, could change this, but has not been tried). Thus the `updateRequired` flag is not used in the kernel because the kernel only runs at all if a rebuild is required; the flag is simply set to zero at the end by the thread handling particle 0.

### VIII. BENCHMARKS AND PERFORMANCE ANALYSIS

To benchmark RUMD we use the Lennard-Jones benchmark described on the LAMMPS homepage, involving an FCC crystal of Lennard-Jones material which is given a kinetic energy sufficient to melt it and then run for  $10^4$  time steps at constant total energy (NVE). Figure 6 shows as black filled symbols the number of timesteps per second (TPS) RUMD can perform on a Gtx780Ti GPU card as a function of system size. For comparison we show also the results published on the LAMMPS homepage for different versions of LAMMPS: A pure CPU version of LAMMPS running on 12 Intel Xeon cores (dual hex-core 3.47 GHz Intel Xeons X5690), and three different GPU-extensions, KOKKOS/CUDA, USER-CUDA, and GPU, all running on a K20x card with 2688 cores (these results are for 100 timesteps). All the GPU-accelerated versions of LAMMPS,

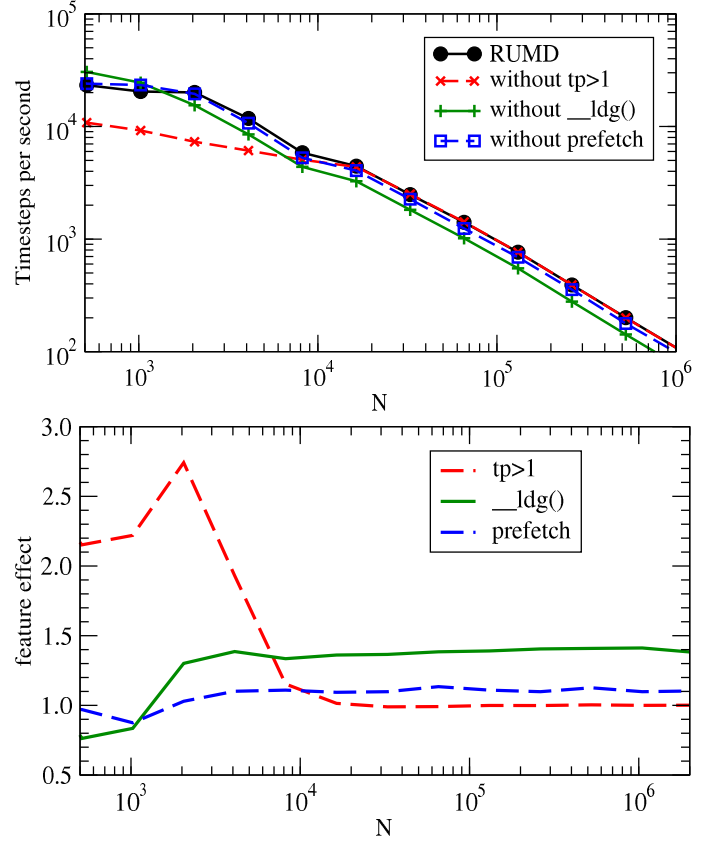


FIG. 7: Analyzing the effect on performance of features of RUMD. The upper panel shows, plotted as in Fig. 6, the performance of the full-RUMD and three other versions in which one feature has been disabled: multiple threads per particle ( $t_p > 1$ ), use of read only data-cache to read positions (`_ldg()`), and pre-fetching. The lower panel shows the same data in terms of the relative boost in performance each feature gives, as a function of system size.

together with RUMD, give similar performance for large  $N$  (above  $\sim 3 \times 10^5$ ). In this regime near perfect scaling with  $N$  is observed, and the GPU versions are 10-20 times faster than LAMMPS running on 12 Intel Xeon cores. Differences show up at small sizes: the number of simulated time steps per second plateaus already at a few tens of thousands of particles for two of the GPU versions of LAMMPS. This plateau means running a simulation with 2000 particles takes as much time as one with 20000 particles; clearly the GPU hardware is under-utilized in this size regime. In fact, for these two implementations (the red and blue curves), it is faster to use the pure CPU version of LAMMPS at the smallest sizes. RUMD, on the other hand, maintains reasonable (though not perfect) scaling down to around  $N = 2000$ . We have included even smaller system sizes in the benchmarking of RUMD, to illustrate that it eventually also begins to plateau - but this only happens when the system size is less than 2000.

Table II gives the parameters chosen by the auto-tuner, as a function of system size. Note that, except for the two smallest system sizes, the auto-tuner chooses the number of threads ( $N \times t_p$ ) to be at least 16000. This illustrates the point made

TABLE II: Performance parameters chosen by the auto-tuner and the resulting TPS (Timesteps Per Second) on a Gtx780Ti card.

N	NB	pb	tp	skin	TPS
512	$N^2$	16	14	0.452	23281
1024	$N^2$	16	10	0.452	20446
2048	$N^2$	48	8	0.5	20068
4096	$N^2$	96	4	0.675	11794
8192	$N$	64	2	0.746	5847
16384	$N$	192	1	0.611	4440
32768	$N$	128	1	0.452	2484
65536	$N$	96	1	0.409	1409
131072	$N$	96	1	0.370	764
262144	$N$	128	1	0.370	390
524288	$N$	128	1	0.370	200
1048576	$N$	128	1	0.335	104
2097152	$N$	96	1	0.335	51

in the introduction, that the number of threads should be much larger than the number of physical cores (here 2880) to get good performance. The reason that fewer threads are used for the two smallest system sizes is probably that the required large  $t_p$  values inflict too large a penalty due to the sequential summation of the  $t_p$  different contributions to the force (see Fig. 3). The switch between the two methods for neighbor-list generation happens at around 8000 particles. In this range of system sizes both methods are sub-optimal and the auto-tuner compensates by increasing the skin size to make neighbor-list updates less frequent.

Figure 7 shows the effect of disabling different optimization features. The upper panel shows the same quantity as in Fig. 6, but with different curves representing different disabled features (the black curve is with all features enabled). The most dramatic difference is when  $t_p = 1$  is enforced, for small and medium systems ( $N < 10^4$ ). No difference is observed at larger  $N$  because there  $t_p = 1$  is the optimal choice, see table II. Disabling the use of the read only data-cache gives the green curve, a significant drop in performance across all sizes except the very smallest  $N < 2000$ , while disabling pre-fetching gives a slight drop, more at larger sizes. The lower panel of Fig. 7 shows the same data, but plotted as the ratio of the speed of the full RUMD to the speed of RUMD with the given feature disabled. Plotting this ratio, on a linear scale, shows the relative effects more clearly. In particular, reading via the read only data-cache gives an effect of order 40%, while pre-fetching has an effect of order 10% at the largest sizes.

## IX. ELECTROSTATICS

A general purpose MD code should include electrostatic (Coulomb) interactions and these should be sufficiently accurate and computationally efficient. While an Ewald-based method which can efficiently handle the long range part of the electrostatic interactions is planned, for our needs so far

N	$2 \cdot 10^3$	$2 \cdot 10^4$	$10^5$
LAMMPS SP	3.26	5.66	6.06
LAMMPS WOLF	2.21	3.57	3.40
LAMMPS PPPM	0.54	0.40	0.35
RUMD SF	9.41	11.3	16.1

TABLE III: Benchmarks for a system of charged Lennard-Jone particles (see text for details) for LAMMPS (CPU) and RUMD (GPU). The metric shown is MATS (millions of atom time steps per second). LAMMPS benchmarks were performed on a Dell 630 server with dual Intel Xeon E5-2699 v3 2.3 GHz CPU's for a total of 36 cores. Coulomb interactions were evaluated using a simple shifted-potential cutoff at distance 6.0 ("SP"), using the Wolf method with the switching parameter  $\alpha$  set to zero ("WOLF", equivalent to the shifted force method), and the Particle-Particle Particle-Mesh method ("PPPM"). RUMD benchmarks were performed on an nVidia GTX 780 Ti, using a shifted force cut-off (SF) at distance 6.0.

we have found it sufficient to use Coulomb forces with a long range shifted-force cut-off as documented in Ref. 50. In that paper it was shown that a shifted-force cutoff of order five inter-particle spacings gives, similar to the Wolf<sup>51</sup> method, results in good agreement with Ewald-based methods in bulk systems.

To benchmark the performance of Coulomb interactions, we performed simulations of a model molten salt in RUMD and the CPU version of LAMMPS. Details of the system simulated are as follows: All particles have identical Lennard-Jones parameters  $\epsilon$  and  $\sigma$ . The charges are  $\pm\sqrt{4\pi\epsilon_0\epsilon\sigma}$  (50% each). The density was  $0.3677 \sigma^{-3}$  and the temperature  $2 \epsilon/k_B$ . The density is the same as was used by Hansen and McDonald in their study of a similar model salt<sup>52</sup>. In Ref.<sup>50</sup> it was shown that a cutoff of  $6.0\sigma$ , corresponding to  $\sim 330$  neighbors per particle at this density, was sufficient to get satisfactory accuracy. The time step is  $0.004 \sqrt{m/\epsilon\sigma}$ .

The data in Table III compare RUMD and the CPU-version of LAMMPS with different methods of evaluating Coulomb interactions and different numbers of CPU-threads. The benchmarks are expressed as MATS (million atom timesteps per second) for ease of comparing different system sizes. The smallest speed-up of RUMD over LAMMPS is here a factor of two. This smaller speed-up compared to the previous section is mostly due to the LAMMPS benchmarks being run on a faster CPU system, a Dell 630 server with 36 XEON cores. For comparison, at the time of writing the cost of the Dell 630 server is roughly 10.000\$, whereas the cost of the GTX 780 Ti card is less than 1000\$ (to this should be added the cost of a fairly standard PC, which can hold three GPU cards).

## X. SUMMARY

We have described the RUMD software package for molecular dynamics simulation on GPUs, concentrating on the optimization strategies that distinguish it from most other GPU MD codes. We have documented its strong performance at small and medium system sizes and performance comparable to other GPU-based MD codes at larger sizes. Work will con-



tinue on RUMD both with regard to features (for example, many-body interactions and efficient long-range Coulomb interactions) and optimization opportunities (for example, dynamic parallelism). The ability to split a simulation over multiple GPUs will also be considered, which will not just allow larger systems (more than the approx. 3 million particles a single card can handle), but also allow even faster simulations of medium systems, given that RUMD already make good use of the hardware for such sizes.

## XI. APPENDIX: THE AUTOTUNER

Here we describe the algorithm used by the autotuner. The basic strategy is to run a series of short simulations (a few hundred to a few thousand time steps) varying the different parameters, to find a set of parameters giving (close to) optimal performance. Not all possible combinations of parameters are attempted in order to reduce the time taken for tuning (for Lennard-Jones-type systems without molecules or Coulomb interactions this is under a minute for small systems, several minutes for larger systems). The initial state of the system is stored so that all comparisons made by the autotuner involve runs of the same length starting from the same configuration.

If the autotuner is not used, RUMD uses default values which depend on the system size: “n2” neighbor-list method for  $N < 8000$ , otherwise “sort”;  $p_b = 32, t_p = 4$  for  $N \leq 10000$ ,  $p_b = 64, t_p = 1$  otherwise. The default skin value is 0.5, which assumes units of length such that the interparticle spacing is of order unity. In principle the default skin should be based on the interparticle spacing (e.g.  $\rho^{-1/3}/2$ ), but in practise length units in MD are generally of order the interparticle spacing and the autotuner can quickly deal with a discrepancy. For some very small systems  $N < 200$  with not too small cutoff it can be faster not to use a neighbor-list at all. The autotuner checks this possibility for systems with  $N < 5000$ .

The dependence of performance on the parameters  $p_b, t_p$  and neighbor-list skin is simple: the time taken shows a single minimum as a function of the parameter. This allows a relatively straightforward optimization strategy to be used. The number of steps run for the different stages depends on the system size (larger for smaller systems sizes to get better timing) and can be altered by the user but should not need to be. The overall strategy is as follows:

1. Run some steps before tuning (default: 10000) to avoid the influence of transient effects (associated for example with having changed the temperature)
2. Run with default parameters to get a baseline performance.
3. Phase 1 optimization: With the default  $p_b$  and  $t_p$  run with the different neighbor-list methods, “none”, “n2”, “sort”. For each one the skin is optimized separately.
4. For the fastest neighbor-list method, and any others within 20% of the fastest, carry out phase 2 optimization: Optimize the parameters  $p_b$  and  $t_p$  using a double

loop: first  $p_b$  considering the values 16, 32, 48, 64, 96, 128 and 192. For each  $p_b$ , values of  $t_p$  are tested starting from 1 and increasing until 64. For each combination of  $p_b$  and  $t_p$  the skin length is re-optimized starting at the last identified optimal value.

5. If the neighbor-list method “n2” is included in phase 2, it can still help to sort the particles once every few hundred times, typically for system sizes near the crossover from “n2” being optimal to “sort” being optimal. This is checked and the optimal sorting interval found.
6. If more than one neighbor-list method was optimized in phase 2, make a final comparison between the phase 2-optimized sorting methods to choose the overall optimized set of parameters (generally, except close to the cross-over from one method to another, the phase 2 optimization does not change which method is chosen, and in that case the difference is small anyway).
7. Run, using the optimized parameters, the same number of steps as for the baseline run to determine the overall improvement due to autotuning.
8. Write the tuned parameters to a file so that repeating the simulation in the same directory with the same “user parameters” does not require re-tuning. For this purpose, “same user parameters” means: same number of particles of each type, same density and temperature and potential parameters (within a tolerance), same integrator type and timestep (within tolerance), and same GPU-type. The actual configuration does not have to be the same. If there is any doubt about re-using the previously found parameters the file can just be deleted.

Some further details are noted here:

- The skin optimization starts from the default value (phase 1) or previously identified phase 1-optimal value (phase 2). Its value is increased and decreased in steps of 20% (phase 1) or 10% (phase 2) until a minimum is identified in the time taken. Attempting to optimize the skin to a precision of better than 10% is not worth the effort.
- The loop over  $t_p$  breaks out when one of the following three conditions is met: (i) the time taken exceeds the minimum time so far three times (ii) the time taken exceeds the minimum time so far by 5% (iii) some GPU resource-limit is exceeded, either the number of threads per block ( $p_b t_p$ ) or the total register count per block.
- The loop over  $p_b$  breaks out when the time taken (having optimized  $t_p$  and skin) exceeds the previous best by 10% or more.
- For very large systems it doesn’t make sense to use anything other than the “sort” method for the neighbor-list. The autotuner omits “none” for  $N > 5000$  and “n2” for  $N > 50000$ . Moreover, large systems generally require larger  $p_b$  and so the autotuner omits  $p_b = 16$  for

$N > 4000$  and  $p_b = 32$  for  $N > 10^5$ . Also for the largest systems only  $t_p = 1$  is relevant; the autotuner omits checking other values for  $N > 10^5$ . These various cutoff-parameters can be changed by the user if (s)he knows their names, but it should not be necessary and this is not included in the documentation.

## Acknowledgments

A large part of this work was sponsored by the Danish National Research Foundation's grant DNRF61.

\* Electronic address: nbailey@ruc.dk

† Electronic address: tbs@ruc.dk

<sup>1</sup> (2014), RUMD software is freely available at <http://rumd.org>.

<sup>2</sup> J. D. Owens, D. Luebke, N. Govindaraju, M. Harris, J. Krueger, A. E. Lefohn, and T. J. Purcell, *Comput. Graph Forum* **26**, 80 (2007).

<sup>3</sup> J. D. Owens, M. Houston, D. Luebke, S. Green, J. E. Stone, and J. C. Phillips, *Proc. IEEE* **96**, 879 (2008).

<sup>4</sup> J. E. Stone, D. J. Hardy, I. S. Ufimtsev, and K. Schulten, *J. Mol. Graphics Modell.* **29**, 116 (2010).

<sup>5</sup> J. Nickolls and W. J. Dally, *IEEE MICRO* **30**, 56 (2010).

<sup>6</sup> R. M. Farber, *J. Mol. Graph. Model.* **30**, 82 (2011).

<sup>7</sup> A. Harju, T. Siro, F. F. Canova, S. Hakala, and T. Ratalaiho (2012), [arXiv:1210.7930](https://arxiv.org/abs/1210.7930).

<sup>8</sup> M. J. Harvey and G. De Fabritiis, *Wiley Interdiscip. Rev.-Comput. Mol. Sci.* **2**, 734 (2012).

<sup>9</sup> J. A. Anderson, C. D. Lorenz, and A. Travesset, *J. Comput. Phys.* **227**, 5342 (2008).

<sup>10</sup> J. A. Anderson and A. Travesset, *Comput. Sci. Eng.* **10**, 8+ (2008).

<sup>11</sup> P. K. Jha, R. Sknepnek, G. I. Guerrero-Garcia, and M. O. de la Cruz, *J. Chem. Theory Comput.* **6**, 3058 (2010).

<sup>12</sup> T. D. Nguyen, C. L. Phillips, J. A. Anderson, and S. C. Glotzer, *Comput. Phys. Commun.* **182**, 2307 (2011).

<sup>13</sup> M. J. Harvey, G. Giupponi, and G. De Fabritiis, *J. Chem. Theory Comput.* **5**, 1632 (2009).

<sup>14</sup> P. Eastman and V. S. Pande, *J. Comput. Chem.* **31**, 1268 (2010).

<sup>15</sup> P. Eastman, M. S. Friedrichs, J. D. Chodera, R. J. Radmer, C. M. Bruns, J. P. Ku, K. A. Beauchamp, T. J. Lane, L.-P. Wang, D. Shukla, et al., *J. Chem. Theory Comput.* **9**, 461 (2013).

<sup>16</sup> P. H. Colberg and F. Höfling, *Comput. Phys. Commun.* **182**, 1120 (2011).

<sup>17</sup> J. Phillips, R. Braun, W. Wang, J. Gumbart, E. Tajkhorshid, E. Villa, C. Chipot, R. Skeel, L. Kale, and K. Schulten, *J. Comput. Chem.* **26**, 1781 (2005).

<sup>18</sup> W. M. Brown, P. Wang, S. J. Plimpton, and A. N. Tharrington, *Comput. Phys. Commun.* **182**, 898 (2011).

<sup>19</sup> R. Salomon-Ferrer, D. A. Case, and R. C. Walker, *WIREs Comput. Mol. Sci.* **3** (2012).

<sup>20</sup> S. Páll, M. J. Abraham, C. Kutzner, B. Hess, and E. Lindahl, *LNCS* **8759**, 3 (2015), proceedings of EASC 2014.

<sup>21</sup> J. E. Stone, J. C. Phillips, P. L. Freddolino, D. J. Hardy, L. G. Trabuco, and K. Schulten, *J. Comput. Chem.* **28**, 2618 (2007).

<sup>22</sup> S. Guo-Liang, W. Jing-Wei, L. Zhen-Hua, W. Wen-Ning, and F. Kang-Nian, *Chem. J. Chinese U.* **29**, 2425 (2008).

<sup>23</sup> W. Liu, B. Schmidt, G. Voss, and W. Müller-Wittig, *Comput. Phys. Commun.* **179**, 634 (2008).

<sup>24</sup> J. A. van Meel, A. Arnold, D. Frenkel, S. F. P. Zwart, and R. G. Belleman, *Mol. Simul.* **34**, 259 (2008).

<sup>25</sup> M. S. Friedrichs, P. Eastman, V. Vaidyanathan, M. Houston, S. Legrand, A. L. Beberg, D. L. Ensign, C. M. Bruns, and V. S. Pande, *J. Comput. Chem.* **30**, 864 (2009).

<sup>26</sup> J. Xu, Y. Ren, W. Ge, X. Yu, X. Yang, and J. Li, *Mol. Simul.* **36**, 1131 (2010).

<sup>27</sup> H. J. Myung, R. Sakamaki, K. J. Oh, T. Narumi, K. Yasuoka, and S. Lee, *Bull. Korean Chem. Soc.* **31**, 3639 (2010).

<sup>28</sup> J. A. Baker and J. D. Hirst, *Mol. Inf.* **30**, 498 (2011).

<sup>29</sup> D. C. Rapaport, *Comput. Phys. Commun.* **182**, 926 (2011).

<sup>30</sup> A. P. Ruymgaart, A. E. Cardenas, and R. Elber, *J. Chem. Theory Comput.* **7**, 3072 (2011).

<sup>31</sup> K. Oguchi, Y. Shibuta, and T. Suzuki, *J. Jpn. I. Met* **76**, 462 (2012).

<sup>32</sup> S. Páll and B. Hess, *Comput. Phys. Commun.* **184**, 2641 (2013).

<sup>33</sup> D. B. Kirk and W.-m. W. Hwu, *Programming Massively Parallel Processors: A Hands-on Approach* (Morgan Kaufmann, 2010).

<sup>34</sup> NVIDIA, *NVIDIA CUDA Compute Unified Device Architecture Programming Guide* (NVIDIA, Santa Clara, CA, USA, 2014), version 6.5.

<sup>35</sup> E. Lindholm, J. Nickolls, S. Oberman, and J. Montrym, *IEEE Micro* **28**, 39 (2008).

<sup>36</sup> NVIDIA, *NVIDIA's next generation CUDA compute architecture: Fermi* (2009), white paper. Available online, Version 1.1, 22 pp.

<sup>37</sup> NVIDIA, *NVIDIA's next generation CUDA compute architecture: Kepler GK110* (2012), white paper. Available online, Version 1.0, 24 pp.

<sup>38</sup> NVIDIA, *NVIDIA GeForce GTX 980, Featuring Maxwell* (2014), white paper. Available online, Version 1.1, 32 pp.

<sup>39</sup> S. Karmakar, C. Dasgupta, and S. Sastry, *PNAS* **106**, 3675 (2009).

<sup>40</sup> Z. Fan, T. Siro, and A. Harju, *Comput. Phys. Commun.* **184**(5), 1414 (2013).

<sup>41</sup> J. Glaser, T. D. Nguyen, J. A. Anderson, P. Lui, F. Spiga, J. A. Millan, D. C. Morse, and S. C. Glotzer, *Comput. Phys. Commun.* **192**, 97 (2015).

<sup>42</sup> S. Toxvaerd and J. C. Dyre, *J. Chem. Phys.* **134**, 081102 (2011).

<sup>43</sup> G. J. Martyna, D. J. Tobias, and M. L. Klein, *J. Chem. Phys.* **101**, 4177 (1994).

<sup>44</sup> T. S. Ingebrigtsen, S. Toxvaerd, O. J. Heilmann, T. B. Schröder, and J. C. Dyre, *J. Chem. Phys.* **135**, 104101 (2011).

<sup>45</sup> T. S. Ingebrigtsen, S. Toxvaerd, T. B. Schröder, and J. C. Dyre, *J. Chem. Phys.* **135**, 104102 (2011).

<sup>46</sup> S. Plimpton, *J. Comp. Phys.* **117**, 1 (1995).

<sup>47</sup> L. Rovigatti, P. Šulc, I. Z. Reguly, and F. Romano, *J. Comp. Chem.* **36**, 1 (2015).

<sup>48</sup> M. P. Allen and D. J. Tildesley, *Computer Simulation of Liquids* (Oxford University Press, 1987).

<sup>49</sup> (2015), thrust Parallel Algorithms Library, URL <http://thrust.github.io/>.

<sup>50</sup> J. S. Hansen, T. B. Schröder, and J. C. Dyre, *J. Phys. Chem. B* **116**, 5738 (2012).

<sup>51</sup> D. Wolf, P. Keblinski, S. R. Phillpot, and J. Eggebrecht, *J. Chem. Phys.* **110**, 8254 (1999).

<sup>52</sup> J. P. Hansen and I. R. McDonald, *Phys. Rev. A* **11**, 2111 (1975).